CHANNEL AND FEATURE CLASSIFIER FUSION IN EEG ANALYSIS

A. Al-Ani
Faculty of Engineering, University of Technology, Sydney
PO Box 123, Broadway, NSW 2007, Australia
e-mail: ahmed@eng.uts.edu.au

Abstract-**This paper investigates the importance of classifier fusion in EEG analysis. Subsets of channels, each represented by a fixed number of features, as well as subsets of individual features from different channels are used to classify EEG trails. The channel/feature subset of each classifier is chosen using a genetic algorithm approach, and the classifier ensembles are formed by finding the appropriate subset of channels/features for each classifier. When applied to the problem of brain-computer interface, the obtained results indicate that the classifier ensemble is able to achieve a higher classification accuracy compared to the best individual classifier.**

***Keywords* - EEG analysis, Classifier combination, channel and feature fusion.**

I. INTRODUCTION

Studies concerned with improving the performance of pattern classification systems have found that different classification algorithms usually offer complementary information about the patterns to be classified [1]. Therefore, if the classification results of different classifiers can be integrated in an efficient way, then the outcome of such combined classifiers can surpass all individual classifiers, including the best one. The basic idea behind combining classifiers is that when making a decision, we should not rely on a single expert (classifier), but rather all experts need to participate in making the decision by combining their individual opinions.

Combining classifiers is recommended when the classifiers are: (i) based on different subsets of features, (ii) based on different classification paradigms, (iii) trained on different subsets of the dataset, and (iv) combination of the above. It is important to mention that better results can only be achieved if the classifiers make different errors and an efficient way that resolves conflicts between the classifiers is found.

As explained in [2], the problem of combining multiple classifiers consists of two parts. The first part, closely dependent on specific applications, includes the problems of "How many and what type of classifiers should be used for a specific application? and for each classifier what type of features should we use?", as well as other problems that are related to the construction of those individual and complementary classifiers. The second part, which is common to various applications, includes the problems related to the question "How to combine the results of different existing classifiers so that a better result can be obtained?" In this work, I will consider problems related to the first issue, and in particular the construction of EEG channel/feature subsets and the fusion of their classification results. This approach aims at improving the analysis of EEG, which is a very challenging problem due to the poor resolution of the signal and its multi-channel nature.

For the particular problem of Brain-Computer Interface (BCI), there has been an extensive research on studying the EEG signals of subjects while performing different mental tasks. Three main aspects can be considered to improve the performance of BCIs, namely signal processing and feature extraction [3], feature and channel selection [4, 5], and classification technique [6]. To the best of my knowledge, the concept of classifier fusion has not been investigated before as an approach to improve the performance of BCIs. In this paper, classifiers will be built using: (i) subsets of channels, where each channel is represented by a fixed number of features, and (ii) subsets of individual features from different channels. These subsets of channels/features will be chosen using a genetic algorithm search technique.

The paper is organized as follows: the next section gives a description of the multiple classifier systems. Section III describes the data and classification method. Description of the proposed EEG classifier ensemble is given in section IV. Section V presents the experimental results, and a conclusion is given in section VI.

II. MULTIPLE CLASSIFIER SYSTEMS

Several methods have been proposed in the literature that aim at improving the performance of pattern classification systems through the combination of multiple classifiers. This is due to the fact that in many applications the optimal classifier[1] does not exist. This is because of two main reasons: (i) in many pattern classification problems a number of classification algorithms can be used. These algorithms are based on different theories and methodologies, and they usually achieve different degrees of success, but the perfection of a particular technique cannot be claimed. (ii) Many types of feature extraction methods are available to represent patterns. In many cases, a concise feature set that preserves the necessary and sufficient information about the patterns to be classified do not exist. In other words, the use of Multiple Classifier Systems (MCS) is motivated by the existence of many alternative solutions to a pattern classification problem, and the observation that these solutions often complement one another in correctness.

There are two broad categories of MCS. The first involves the design of a classifier ensemble, while the other deals with developing a method to combine classifiers' outputs. A good review of MCS is given in [7].

The design of classifier ensemble, which is the focus of this paper, can be achieved using one or more of the following:

[1] The following definition is adopted for the optimal classifier: if a pattern is misclassified by the optimal classifier, then no other classifier can correctly classify that pattern.

- Dividing the training set into smaller subsets that will be used to train a number of classifiers. Hence, the aim is to achieve complementary results by training the classifiers with different subsets. Examples of this approach are the bagging and boosting [8].
- Different classifier settings. This includes different classifier architectures and different initialization of the classifier's learning parameters. For example, a multi-layer Perceptron neural network can be designed using different number of hidden layers and/or units in each layer. Also, different weight initialization can be used.
- Different classifier types, such as: Baysian classifier, multi-layer Perceptron neural network, and a k-nearest neighbor classifier, etc.
- Different features to train the classifiers. The original feature set is divided into smaller subsets, and hence, complementary classifiers are trained with different feature subsets. Example of this approach is the random subspace method [9].

Among those approaches, I will be focusing on the last one and expand it to accommodate the multi-channel nature of EEG. It is important to mention that different methods have been proposed in the literature to split the feature space into smaller subspaces. Some of the famous methods are: random selection [9], genetic algorithm [10] and forward/backward selection [11]. Because the Genetic Algorithm (GA) has proven to be a powerful search method to a number of applications, it will be used here to select the channel/feature subsets of each classifier.

III. Description of Data and Classification Method

The data used here was obtained from the Department of Medical Informatics, University of Technology, Graz, Austria[2]. EEG signals were recorded for three right handed females with 56 Ag/AgCl Electrodes using monopolar montage, with reference electrode on the right ear. Fig. 1 illustrates the position of electrodes. Each subject was placed in an armchair and asked to imagine right and left finger movements according to stimuli on screen. A total of 8 seconds of data were recorded at 128 Hz sampling rate, 2 seconds before the stimulus and 6 after it. A total of 406 trials were used, 208 for the left movement and 198 for right movement. More details on experiment set-up can be found in [12].

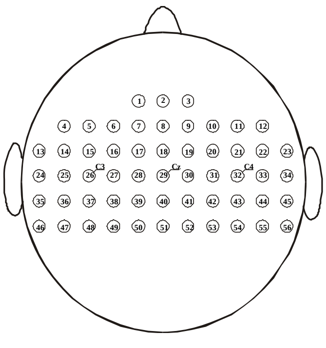


Fig. 1. Position of EEG electrodes

Since EEG is a time-varying and space-varying non-stationary signal, we used the wavelet transform to extract features from the data of each trial. According to [13], the wavelet transform was found to provide good way to visualize and decompose EEG signals into measurable component events. The decomposition of the signal using this tool gives scaled and shifted version of the original "mother" wavelet. The first order Daubechi (db1) was chosen as the mother wavelet in this work. A subset of 3 features that represent the energy values of the frequency bands 4-8, 8-16 and 16-24 Hz was found to give a good compromise between performance and the use of small number of features to represent each channel [14].

A Bayesian classifier will be used to classify the extracted features into one of the two classes. The Bayesian classifier[3] estimates the posterior probability of each class, which is more useful in combining classification results than classifiers that produce abstract (or binary outputs). Also, because of the limited number of training patterns, other classification algorithms, such as multi-layer perceptron neural networks may not be suitable. To generate the Bayesian classifier, a greedy expectation maximization algorithm is used to estimate the parameters of a Gaussian model (the original expectation maximization algorithm is not suitable because of the insufficient amount of training data).

IV. The Proposed Classifier Ensemble

An important part in the design of classifier ensemble that uses different features to train the classifiers is the choice of appropriate set of features for each classifier. As mentioned earlier, this can be achieved using a search algorithm such as the forward (backward) search and a GA-based search procedure. The selection of features can also be implemented using heuristic methods or diversity measures [7].

Because of the multi-channel nature of EEG (the additional spatial dimension), two scenarios will be investigated. In the first scenario, a channel ensemble will be constructed, where each channel is represented by the three features described in the previous section, and each classifier, which will be referred to as "*channel classifier*", will use a subset of channels as its input. On the other hand, individual features from different channels will be used to construct the ensemble in the second scenario, i.e., a collection of "*channel/feature classifiers*". The GA will be used to select the channels/features of each classifier in both cases. To construct the different classifiers, two approaches will be examined. In the first approach, a heuristic method will be used, which aims at maximizing the classification accuracy of each classifier (by selecting certain channels/features) without considering any sort of collaboration between the

[2] The author would like to thank the Department of Medical Informatics, University of Technology, Graz, Austria for providing the data.

[3] The GMMBayes Matlab Toolbox is used here (http://www.it.lut.fi/project/gmmbayes).

different classifiers. The obtained classification results will then be averaged to calculate the classification accuracy of the ensemble. In the other approach, the classification result of the ensemble will be calculated by averaging the posterior probability of the different classifiers, and the aim will be the maximization of the ensemble classification accuracy. In order to reduce the search complexity, classifiers will be formed one at a time instead of constructing the whole ensemble at the same time.

The GA, which is a combinatorial search technique based on both random and probabilistic measures, was chosen to search through the channel/feature spaces because it has proven to be a powerful optimization method. Subsets of channels/features are evaluated using a fitness function and then combined via cross-over and mutation operators to produce the next generation of subsets. The GA employs a population of competing solutions that evolve over time, to converge to an optimal solution. Effectively, the solution space is searched in parallel, which helps in avoiding local optima. A GA-based variable selection solution would typically be a fixed length binary string representing a channel/feature subset, where the value of each position in the string represents the presence or absence of a particular channel/feature. In this work, a GA-based selection method is implemented using the average classification accuracy of a ten-fold cross-validation as the fitness function. Trials from all three subjects are used, i.e., the experiments are not subject dependent.

V. EXPERIMENTAL RESULTS

In order to compare between the performance of channel and channel/feature subsets, the GA-based selection method described earlier was used to select subsets of features that maximize the classification accuracy. The desired number of channels (or individual features/3) was varied between 2 and 16. The obtained results, shown in Fig. 2, indicate that in most cases the channel/feature classifiers outperform channel classifiers. This is expected as the formation of channel subsets represents a special case of individual feature subsets. However, the searching space of channel/feature subsets is more complex than that of the channel subsets. This becomes more apparent when the number of all possible subsets gets bigger, which explains the small difference in performance between the two methods in some cases. The figure also shows that the difference in performance between the training and test sets increases as the number of channels/features increases, i.e., the input attributes to the classifier. This is mainly due to the limited number of the training patterns.

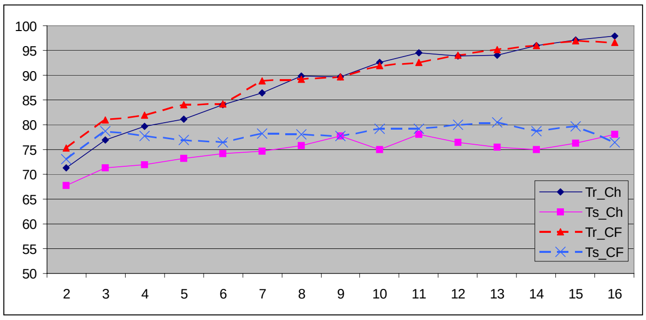
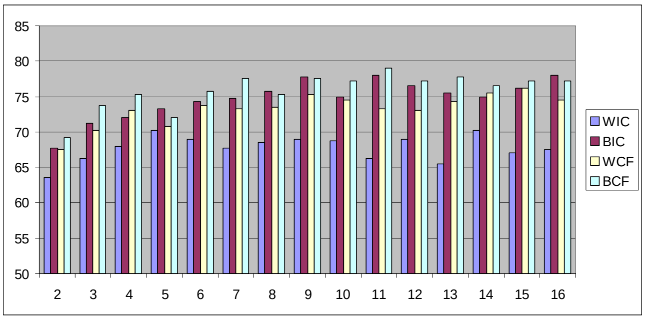
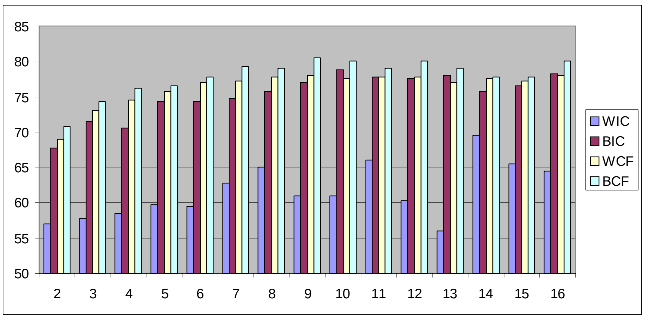


Fig. 2. Classification accuracy of training and test sets for both channel and channel/feature selection

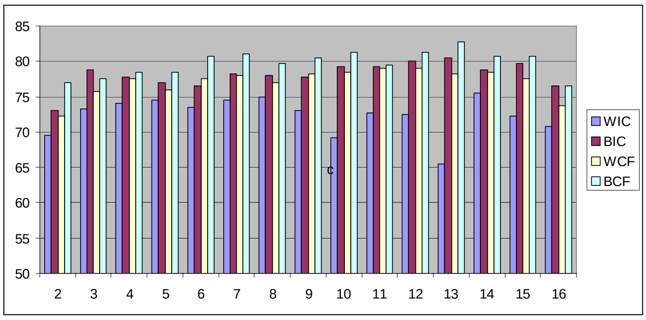
In the second experiment, two methods have been adopted to construct classifier ensembles, as explained in the previous section. The heuristic method independently maximizes the classification accuracy of each classifier, while the second method maximizes the classification accuracy of the ensemble. The classifier ensembles are formed using 5, 4, and 3 classifiers, considering the desired number of channels to range between [2, 10], [11, 13], and [14, 16] respectively. Fig. 3 shows the performance of the Worst Individual



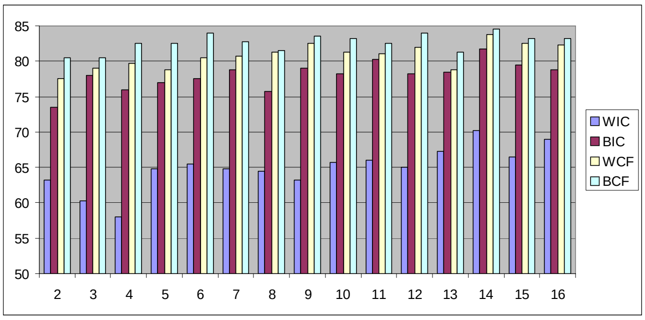
(a) Heuristic channel fusion



(b) Maximized average channel fusion



(c) Heuristic channel/feature fusion



(d) Maximized average channel/feature fusion

Fig. 3. Classification accuracy of WIC, BIC, WCF and BCF using different no. of channels/features. 2, 3, 4 and 5 classifiers (C2…C5) are considered

Classifier (WIC), Best Individual Classifier (BIC), Worst Classifier Fusion (WCF), and Best Classifier Fusion (BCF) using the two methods for both channel and channel/feature classifiers. The results indicate that the heuristic method is not a powerful way to construct classifier ensembles, where in many cases the classification accuracy of the WCF is found to be less than that of the BIC, and in certain cases the BCF performs worse than the BIC. Maximizing the classifier ensemble on the other hand gives quite good results, where in all cases the BCF performs better than the BIC, and only in few cases the BIC outperforms the WCF. In spite of that, the figure shows that the performance of the WIC for the heuristic measure is clearly better than that of the maximized average (the worst case for the channel fusion is 63.5% and 56% respectively). This emphasis the importance of collaboration between the different classifiers, as selecting good individual classifiers does not always lead to the construction of good ensembles. Finally, similar to the first experiment, this experiment also confirms the superiority of channel/feature selection over channel selection, where the best classification accuracy of the fused channel/feature classifiers is found to be 84.5% compared to 80.5% for the channel classifiers. If we define the Error Reduction Rate (*ERR*) as:

$$ERR = \frac{BCF - BIC}{100 - BIC} \times 100\% \qquad (1)$$

then the best *ERR* for both channel and channel/feature fusion would be 19.49% and 28.89% respectively.

VI. Conclusion

The importance of channel and channel/feature classifier fusion has been investigated. Experiments conducted using the problem of brain computer interface confirmed that the fusion of a number of classifiers outperformed the best individual classifier. It has also been found that the construction of classifier ensemble using individual features from different EEG channels could usually achieve better results than using a fixed number of features to represent each channel. The experiment also proved the importance of collaboration between classifiers in achieving good results, as the construction of classifier ensembles using a number of good individual classifiers has not guaranteed a good performance.

References

[1] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12 pp 993-1001, 1990.

[2] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. Systems, Man and Cybernetics*, 22:418_435, 1992.

[3] P. Sykacek, S. Roberts, M. Stokes, E. Curran, M. Gibbs and L. Pickup, "Probabilistic Methods in BCI Research," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 11, 2003, pp 192-195.

[4] M. Pregenzer and G. Pfurtscheller, "Frequency component selection for an EEG-based brain to computer interface," *IEEE Trans. on Rehabilitation Engineering*, vol. 7, 1999, pp 413-419.

[5] T.N. Lal, M. Schrder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer and B. Schlkopf, "Support Vector Channel Selection in BCI," *IEEE Trans. on Biomedical Engineering*, vol. 51, 2004, pp 1003-1010.

[6] R. Palaniappan, R. Paramesran, S. Nishida and N. Saiwaki, "A New Brain Computer Interface Design Using Fuzzy ARTMAP", *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 10, 2002, pp 140-142.

[7] F. Roli and G. Giacinto, "Design of multiple classifier systems," In H. Bunke and A. Kandel, editors, *Hybrid Methods in Pattern Recognition*, World Scientific Publishing 2002, pp 199–226.

[8] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization," *Machine Learning*, vol. 40, 1999, pp 139-158.

[9] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, 1998, pp 832-844.

[10] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Feature Selection for Ensembles: A Hierarchical Multi-Objective Genetic Algorithm Approach", *In Proceedings of 7th Intl. Conf. on Document Analysis and Recognition*, 2003, pp 676- 680.

[11] A. Tsymbal, P. Cunningham, M. Pechenizkiy and S. Puuronen, "Search Strategies for Ensemble Feature Selection in Medical Diagnostics," Technical report, Trinity College Dublin, 2003.

[12] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. on Rehabilitation Engineering*, vol. 8, 2000, pp. 441-446.

[13] V.J. Samar, "Wavelet Analysis of Neuroelectric Waveforms", *Brain and language*, vol 66, 1999, pp 1-6.

[14] A. Al-Ani and A. Al-Sukkar, "Effect of feature and channel selection on EEG classification," *IEEE Intl. Conference of the Engineering in Medicine and Biology Society* (*EMBC 2006*) – to appear.